Integrative Classification and Clustering Using Cancer Genomics Data

Obulkasim, A.

2015

document version
Publisher's PDF, also known as Version of record

Appendix A

This appendix provides supplementary information of a toy example for calculation of reclassification score and results from our stepwise classifier using different algorithm combinations not showed in the Chapter 2.

Reclassification score calculation with toy example

Let say we have 10 samples in training sets (both in clinical and molecular data). After applying selected classification algorithm to the two data types separably and compare them with true class labels we have following:

$$Y.cli = (1, 0, 1, 1, 0, 0, 0, 0, 0, 1), \;\; Y.gen = (1, 1, 1, 1, 0, 1, 1, 0, 1, 1),$$

where $Y.cli$ denotes the classification result from clinical data and $Y.gen$ from molecular data. 0 means sample is wrongly classified and 1 means correctly classified. Based on above result, we set $K = 2$. Project a test sample (i) onto the clinical data space and measure proximity between each of those training samples. Let say after ordering (descending) the proximity values, it produces following order index:

$$order.index^{cli} = (10, 3, 6, 4, 1, 9, 5, 2, 7, 8),$$

use this order index to re-order the $Y.cli$

$$Y.cli.ordered = (1, 1, 0, 1, 1, 0, 0, 0, 0, 0)$$

Let's take $K = 2$ nearest neighbor from the correctly and the incorrectly classified samples groups, separately. Weighted rank will be

$$C^R_{i1} = 1 \times \frac{1}{1} = 1, \;\; C^R_{i2} = 2 \times \frac{1}{2} = 1$$

In similar way we calculate

$$C^W_{i1} = 3 \times \frac{1}{1} = 3, \;\; C^W_{i2} = 6 \times \frac{1}{2} = 3.$$

Sample indexes of closet correctly classified $K = 2$ neighbors of the test sample (i) are

$$_{CR(i1)} = 10, \;\; _{CR(i2)} = 3$$

indexes for the incorrectly classified group are

$$_{CW(i1)} = 6, \;\; _{CW(i2)} = 9.$$

Now, we gain two group of samples indexes (for correctly and incorrectly classified) of training samples which are close to the test sample (i). In next step, based on these sample indexes we search for the nearest neighbors for them in the molecular

data space one by one. Let's consider $_{CR(i1)} = 10$. Let say after ordering (descending) the proximity values with respect to 10^{th} sample, it produces following order index:

$$order.index^{gen} = (7, 5, 8, 4, 2, 3, 1, 10, 6, 9),$$

use this order index to re-order the $Y.gen$

$$Y.gen.ordered = (1, 0, 1, 1, 1, 1, 1, 1, 1, 0)$$

based on this we calculate

$$G^R_{10} = 1\times\frac{1}{1}+3\times\frac{1}{2} = 2.5,\;\; G^W_{10} = 2\times\frac{1}{1}+10\times\frac{1}{2} = 7,\;\; G_{10}(G_{CR(i1)}) = 7-2.5 = 4.5.$$

Let say in similar way we calculate

$$G_3(G_{CR(i2)}) = 3,\;\; G_6(G_{CW(i1)}) = 1.5,\;\; G_9(G_{CW(i2)}) = 5.$$

Final aggregated information for the i^{th} test sample will be

$$Right_i = 1 \times 4.5 + 1 \times 3 = 7.5, \quad Wrong_i = 3 \times 1.5 + 3 \times 5 = 19.5.$$

As a result, reclassification score for this test sample is

$$RS_i = 7.5 - 19.5$$

Results from different algorithm combinations

First we present the case where the clinical data performs better than the molecular data using prostate cancer data. Then, for the sake of completeness, we show results from various combination of classifier.
The prostate cancer data [Stephenson et al., 2005] contains 79 samples, 37 with and 42 without recurrent primary prostate tumors. Pre-filtered gene expression data contains 7884 genes and the clinical factors are composed of serum PSA level (nominal), Gleason stage (ordinal), extra capsular extension (nominal), surgical margin (binary), seminal vesicle invasion (binary), lymph node involvement (binary), TNM (nominal), age (nominal).

The case where clinical data performs better than molecular data. Figure 9-10 illustrates the accuracy of the stepwise approach for the prostate cancer data. We use the RF for clinical data and the Plsrf-x for molecular data. The IntegrativeME method with sPLS feature selection attain the highest accuracy (76%). The accuracy from the stepwise approach is somewhat lower than the one from the IntegrativeME, keeping in mind that IntegrativeME requires 100% molecular data to achieve this. Next, we apply Plsrf-x-pv to molecular data. Since we do not have the classification result with this algorithm from the IntegrativeME, we only compare with the result from the Plsrf-xz-pv [Boulesteix et al., 2008]. As we observe from the Figure 9-10, result from Plsrf-xz-pv is almost the same as from clinical data. The stepwise approach reaches its climax at the beginning as it should and its accuracy is comparative.

In the following part, we present the results of the stepwise classification approach on three data sets with different algorithm combinations.

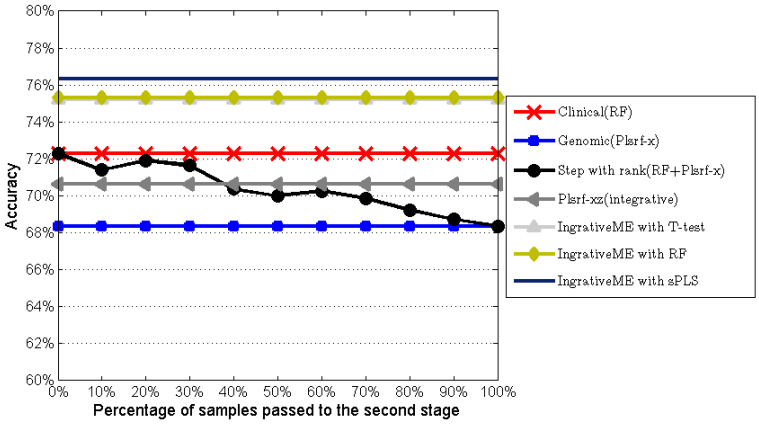

Figure 9: Illustration of the scenario where clinical data is preferred over molecular data using the prostate cancer data. Here, the `RF` *classifier is applied to clinical data, and the* `Plsrf-x` *classifier is applied to expression data.*

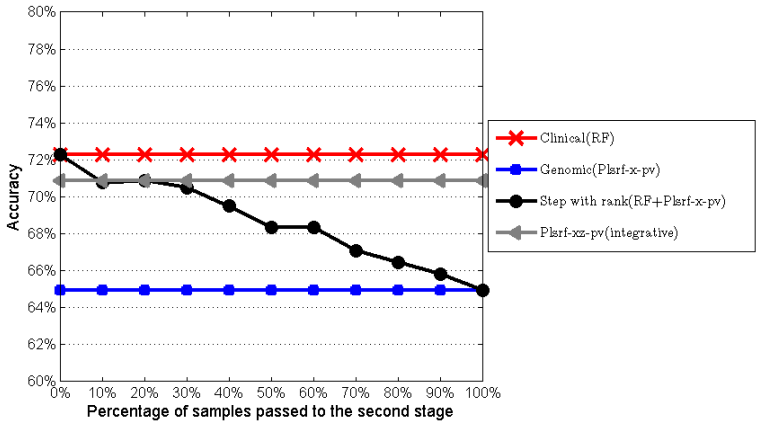

Figure 10: Illustration of the scenario where clinical data is preferred over molecular data using the prostate cancer data. Here, the `RF` *classifier is applied to clinical data, and the* `Plsrf-x-pv` *classifier is applied to expression data.*

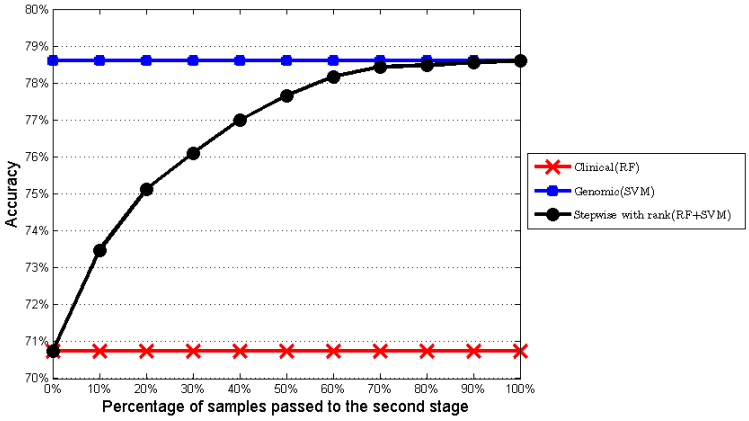

Figure 11: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data. Here, the `RF` *classifier is applied to clinical data, and the* `SVM` *classifier is applied to expression data.*

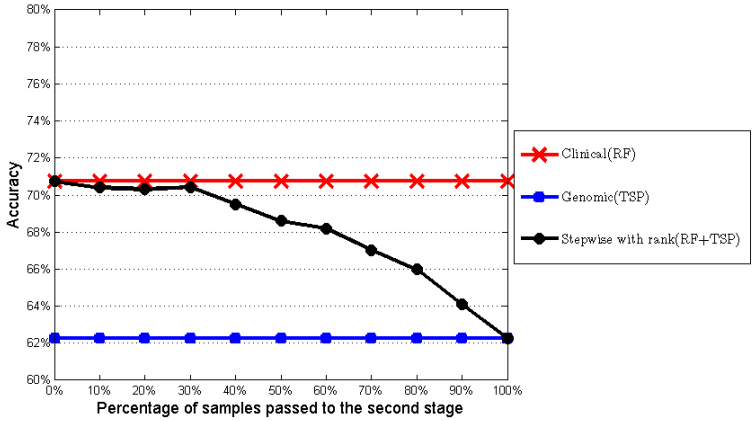

Figure 12: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data. Here, the `RF` *classifier is applied to clinical data, and the* `TSP` *classifier is applied to expression data.*

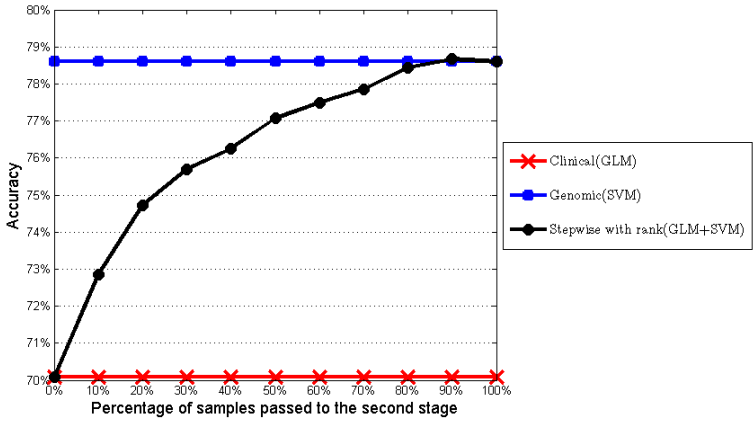

Figure 13: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data, Here, the `GLM` *classifier is applied to clinical data, and the* `SVM` *classifier is applied to expression data.*

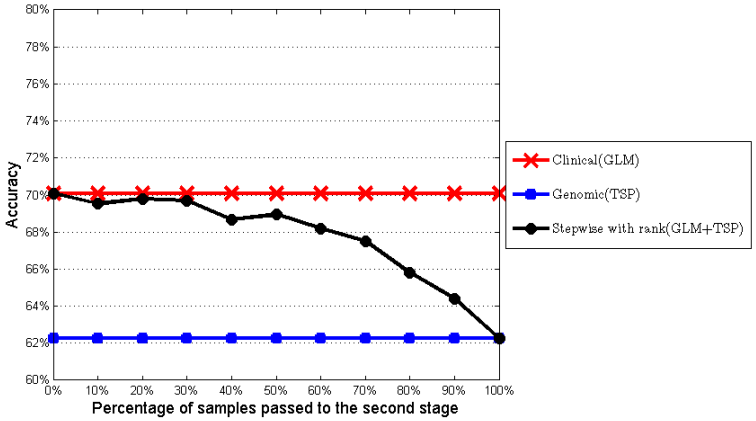

Figure 14: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data, Here, the `GLM` *classifier is applied to clinical data, and the* `TSP` *classifier is applied to expression data.*

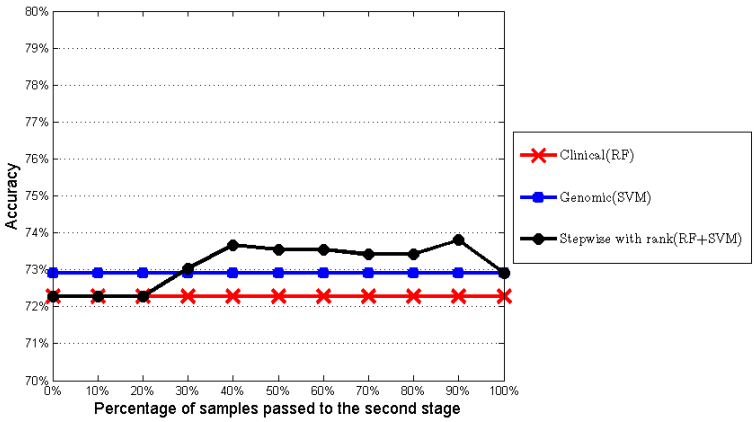

Figure 15: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data. Here, the `RF` *classifier is applied to clinical data, and the* `SVM` *classifier is applied to expression data.*

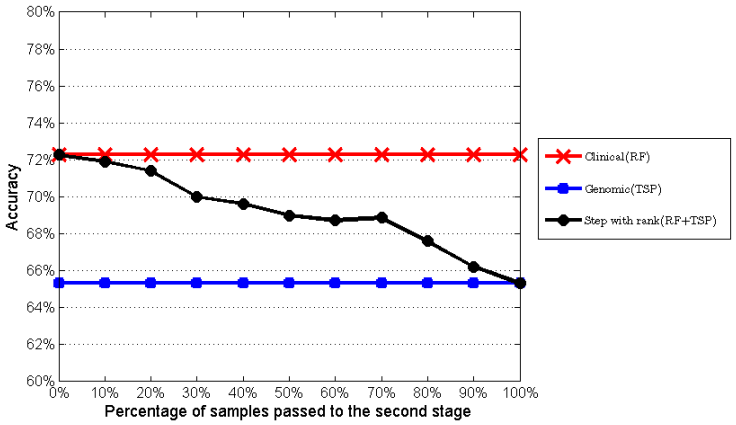

Figure 16: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data. Here, the `RF` *classifier is applied to clinical data, and the* `TSP` *classifier is applied to expression data.*

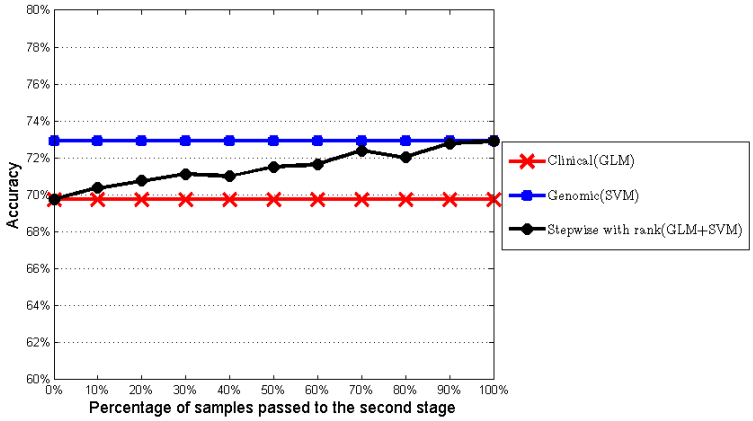

Figure 17: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data, Here, the `GLM` *classifier is applied to clinical data, and the* `SVM` *classifier is applied to expression data.*

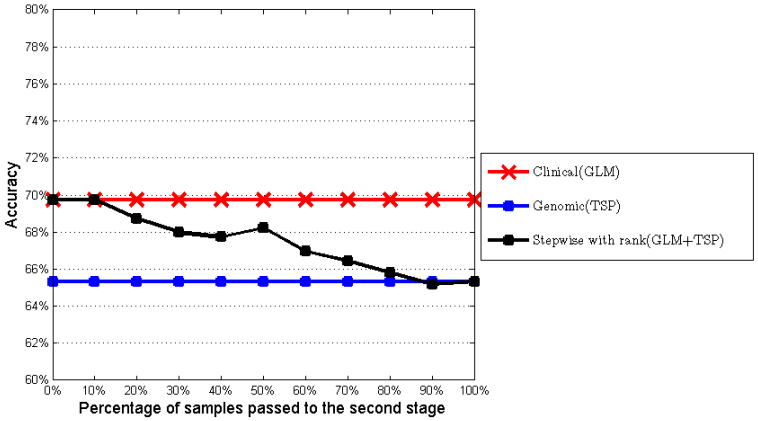

Figure 18: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data, Here, the `GLM` *classifier is applied to clinical data, and the* `TSP` *classifier is applied to expression data.*

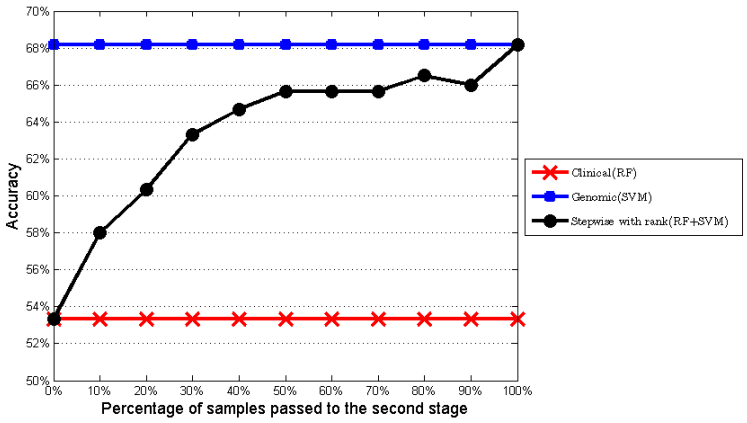

Figure 19: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data. Here, the `RF` *classifier is applied to clinical data, and the* `SVM` *classifier is applied to expression data.*

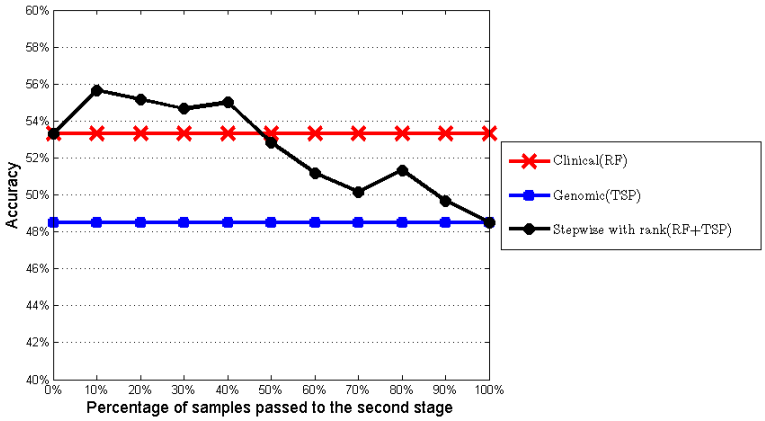

Figure 20: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data. Here, the `RF` *classifier is applied to clinical data, and the* `TSP` *classifier is applied to expression data.*

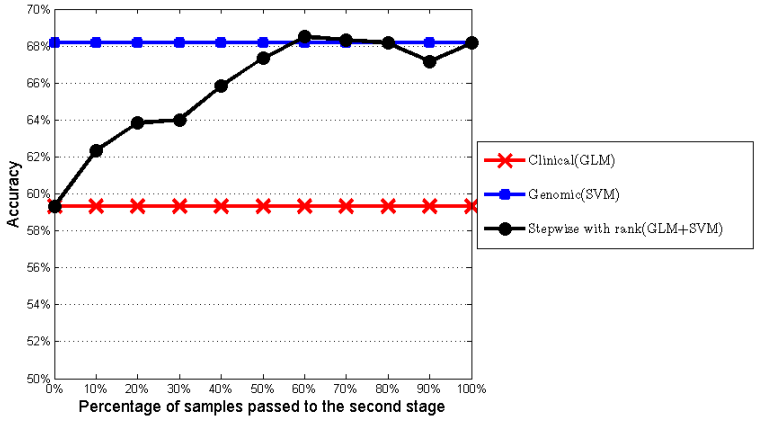

Figure 21: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data, Here, the `GLM` *classifier is applied to clinical data, and the* `SVM` *classifier is applied to expression data.*

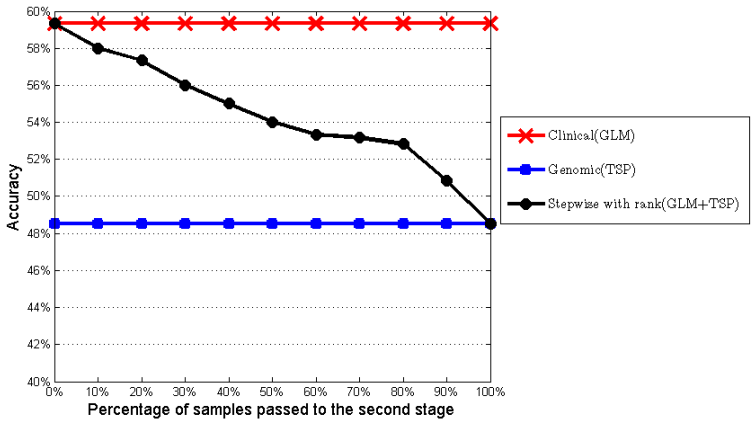

Figure 22: Illustration of the scenario where clinical data is preferred over molecular data using the breast cancer data, Here, the `GLM` *classifier is applied to clinical data, and the* `TSP` *classifier is applied to expression data.*

Appendix B

This appendix provides supplementary information not mentioned in Chapter 4.

Motivational example

Sørlie et al. [2001] used gene expression data to cluster the breast carcinomas using hierarchical clustering and correlate the extracted clusters to the clinical outcome. They reported six clusters, each of them with unique clinical characteristics (Figure 23). After closely examining the HC tree we find that, the reported six clusters cannot be extracted by cutting with a straight line at any place. At most, five out of the six clusters (solid line) can be retrieved unless one introduces a piecewise cut to the left branch (broken line). Observe that, although the Basel-like and the ERBB2+ tumors are very different in terms of molecular (i.e. TP53 status) and clinical (i.e. survival time) features, the fixed-height cut approach fails to separate them. As this real life example shows, sometimes the informative clusters on a branch of the HC tree are located at a deeper level than the other branches.

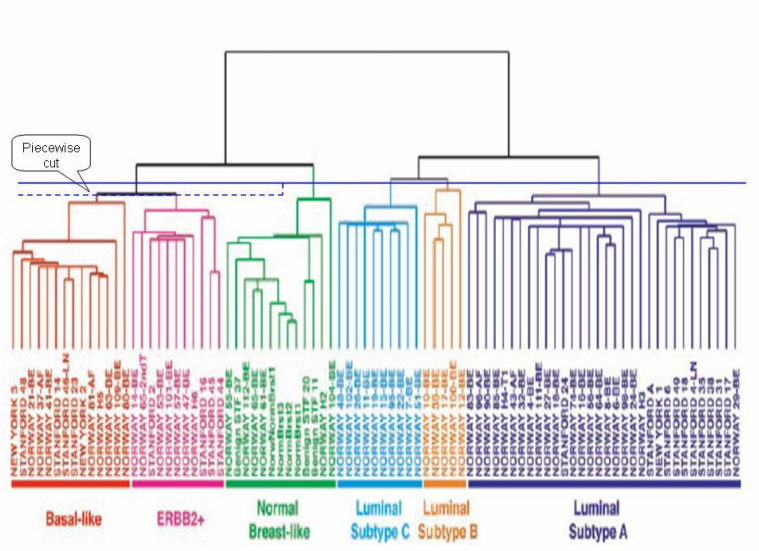

Figure 23: The piecewise cut versus the straight-line cut. The HC tree is derided from using the gene expression data by Sørlie et al. [2001].

Differences between the piecewise cut and existing approach

The piecewise cut approach we proposed here differs from the HC clustering method proposed by Dotan-Cohen et al. [2007] in the following aspects

- The latter approach is designed for gene clustering; application to clustering samples (tissues/disease/patients) is not straightforward. As emphasized by Baya and Granitto [2011], clustering samples is very different from clustering genes. On the other hand, our approach is designed for clustering samples using all types of genomic data.
- The approach by Dotan-Cohen et al. [2007] cuts HC tree at any place as long as resulting clusters are maximally consistent with pre-defined gene labels. This implies that the HC tree structure is allowed to change dramatically. We believe that one HC tree expresses the tendency of observations to be clustered by their signatures in the data from which HC tree is derived. Although returned clusters are homogeneous in terms of partially available observation labels, this comes at the cost of losing its "honesty" to the data set HC is based upon. To preserve the HC tree structure, our approach selects cuts so that only the observations which are neighbors in the leaf nodes form clusters. For example in Figure 4.4, no sample in the *piecewise.cluster1* is not allowed to form a cluster with samples in *piecewise.cluster3*.
- The approach by Dotan-Cohen et al. [2007] requires discretized background information; our approach, however, does not has any restriction on the format of available background information. To the best of our knowledge, we are the first one to implement a HC tree snipping scheme which utilizes commonly available, clinically most relevant patient follow-up data as background information.

Comparison with the fixe-height cut approach: results not included in Chapter 4

Note that, except for the C-index, we use the R package `fpc` [Hennig, 2010] to calculate WSS and GK measures. For a given partition and corresponding distance matrix, *cluster.stats* function in this package returns a overall quality score for it. Clustering performances from the piecewise and the fixed-height approaches are given in Table 5-6.

Association between the optimal clusters found on the Lung.1 data set and the external clinical outcomes

In our analysis, we used the patient follow-up information in the cluster extraction process. To make the comparison between the piecewise and the fixed-height cut methods thoroughly enough, we go one step further to check the association between the optimal partitions retrieved by two approaches and the clinical outcomes not used in the cluster extraction process. For this illustration we used the Lung.1 data set [Beer et al., 2002].

Besides the follow-up data, other clinical information such as disease stage and differentiation were also available for this data set. Chi-square test is used to check the association between the aforementioned two clinical variables and the the best partitions found when the different quality measures are used. Comparison results are given in Table 4.

Regardless of which partition evaluation criteria is used, the fixed-height cut approach generates the same result (a partition with two clusters). The piecewise approach on the other hand, produces slightly different results in different criteria settings. In all cases, the best partitions selected by the fixed-height approach exhibits no association with Stage or Differentiation. While the optimal partitions from the piecewise approach exhibit weak associations with Differentiation, strong associations with Stage are observed. These results further show the potential

Table 2: Comparison of the error rates from the two approaches when the C-index is used for the gene expression data. In each column, numbers denote the number of times our method produces smaller error rates than the fixed-height cut in 100 repetitions, and the opposite holds for numbers in parentheses.

Data	**Ward**		**PNN+Concordance**	
	AIC	BIC	AIC	BIC
Lung.1	61(32)	61(33)	15(8)	15(9)
Leukemia	67(33)	73(27)	81(19)	91(7)
Lung.2	65(35)	62(37)	74(22)	74(20)
Lymphoma	57(43)	66(34)	68(31)	82(18)
Prostate	45(54)	51(49)	65(34)	72(28)
Glioblastoma	67(33)	73(27)	54(44)	65(35)

Table 3: Comparison of the error rates from the two approaches when the GK is used for the gene expression data. In each column, numbers denote the number of times our method produces smaller error rates than the fixed-height cut in 100 repetitions, and the opposite holds for numbers in parentheses.

Data	**Ward**		**PNN+Concordance**	
	AIC	BIC	AIC	BIC
Lung.1	63(31)	63(30)	14(12)	15(16)
Leukemia	71(27)	70(28)	74(24)	79(16)
Lung.2	53(47)	51(49)	74(20)	68(24)
Lymphoma	61(38)	62(37)	65(32)	79(20)
Prostate	44(49)	42(41)	52(39)	53(29)
Glioblastoma	46(42)	52(19)	45(42)	42(28)

Table 4: Association between the optimal partitions generated by the piecewise and the fixed-height (in parentheses) cuts with the external clinical outcomes.

Method		**Stage**	**Differentiation**
AIC with	WSS	0.052(0.859)	0.348(1)
	C-index	0.051(0.859)	0.434(1)
	GK	0.002(0.859)	0.674(1)
BIC with	WSS	0.052(0.924)	0.373(1)
	C-index	0.053(0.924)	0.517(1)
	GK	0.002(0.924)	0.690(1)

of the piecewise cut approach. Note that the p-values we present here are bias, because we did not re-snip the HC tree under permutation. The potential bias caused by the follow-up information which we used during the cluster extraction process is correlated with the two new clinical outcomes we tested here. Here, we only want to show the differences in the magnitude of the p-values from the two approaches.

Visualization of the cuts induced by the two approaches and the survival curves of the resulting clusters

Here, we present HC trees and the cuts induced by the two approaches not included in Chapter 4. In all illustrations the BIC criteria is used for follow-up data, and the KG criteria is used for expression data. In each HC tree, the blue broken line denotes the cut induced by the fixed-height cut approach, and the red rectangles corresponds to the cuts from the piecewise cut approach. The number in each leaf node denotes the survival time and the event status indicator, respectively.

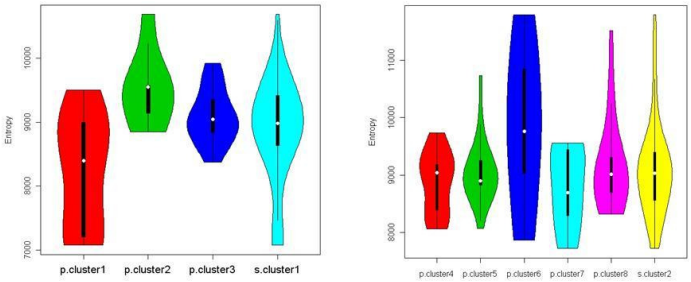

Figure 24: Left figure (violin plot) shows the entropy distributions correspond to clusters on the left branch of the HC tree in Figure 4.4 derived from the Leukemia data set [Bullinger et al., 2004]. Right figure corresponds to clusters on the left branch.

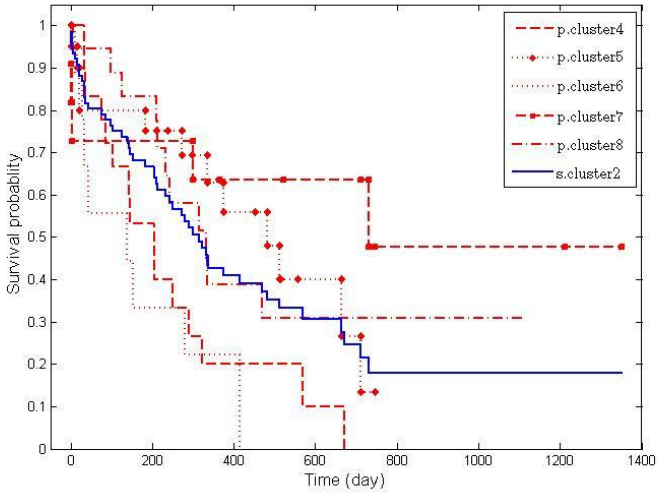

Figure 25: The survival curves correspond to clusters on the right branch of the HC tree in Figure 4.4.

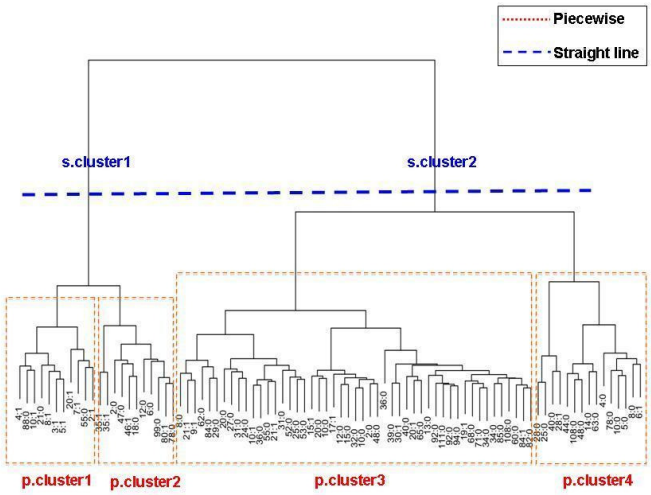

Figure 26: The HC tree corresponds to the Leukemia data set [Beer et al., 2002], and the optimal cuts induced by the two approaches.

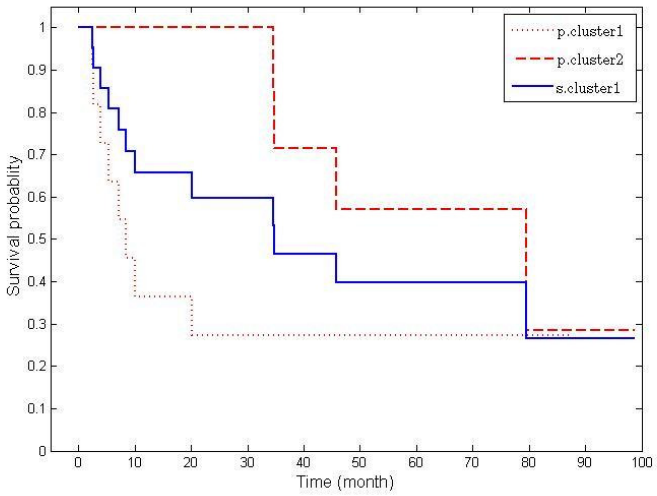

Figure 27: Kaplan-Meier survival curves correspond to the clusters on the left branch of the HC tree in Figure 26.

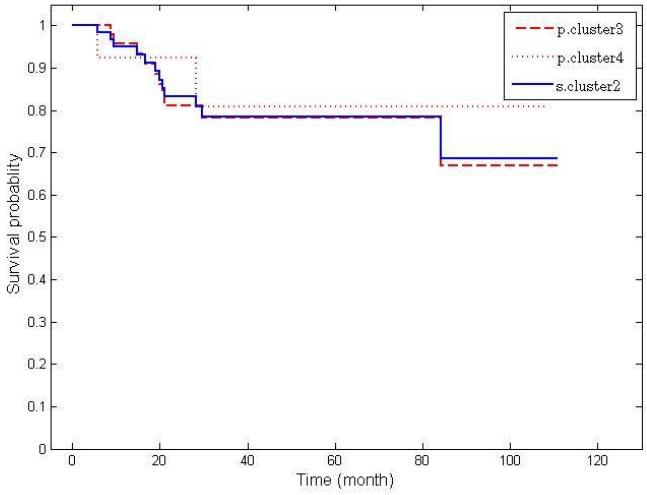

Figure 28: Kaplan-Meier survival curves correspond to the clusters on the right branch of the HC tree in Figure 26.

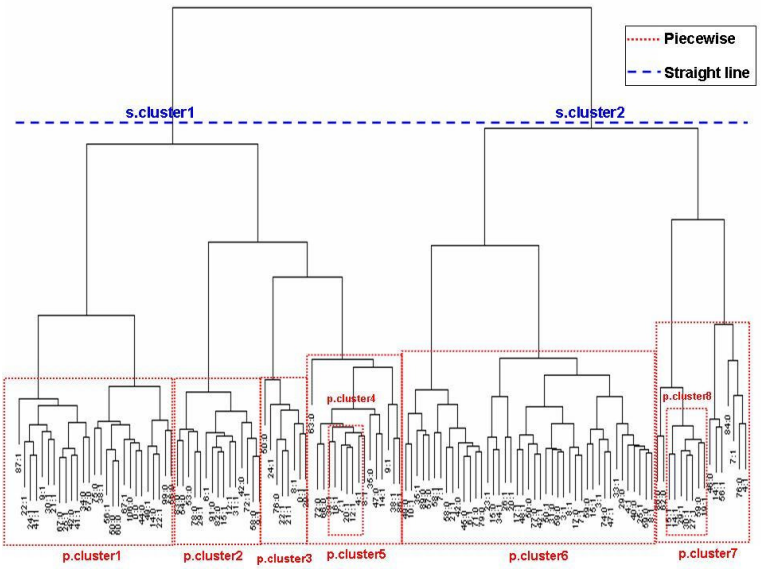

Figure 29: The HC tree corresponds to the Lung.2 data set [Bhattacharjee et al., 2001], and the optimal cuts induced by the two approaches.

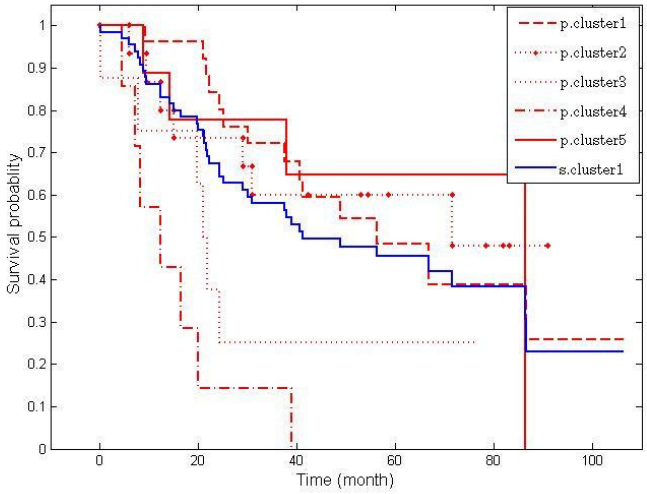

Figure 30: Kaplan-Meier survival curves correspond to the clusters on the left branch of the HC tree in Figure 29.

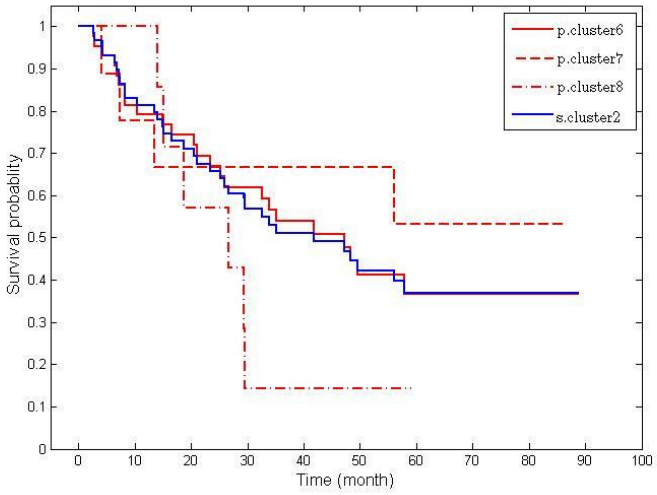

Figure 31: Kaplan-Meier survival curves correspond to the clusters on the right branch of the HC tree in Figure 29.

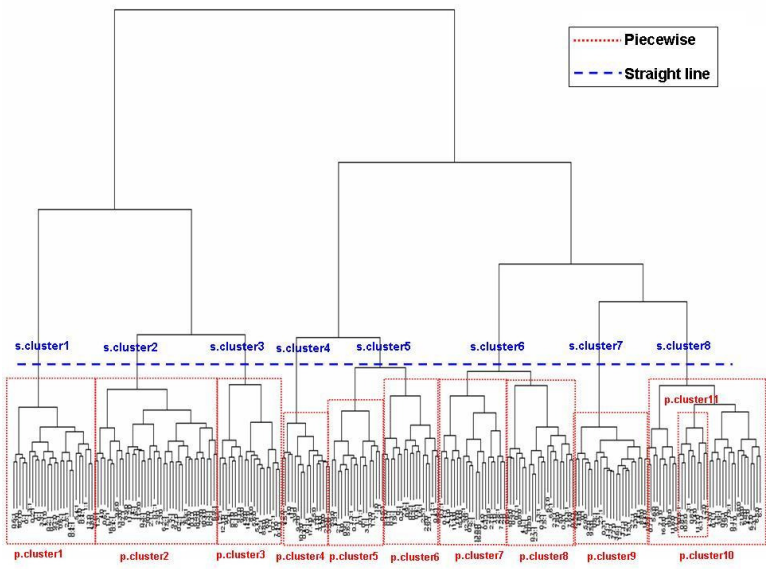

Figure 32: The HC tree corresponds to the the Lymphoma data set [Rosenwald et al., 2002], and the optimal cuts induced by the two approaches.

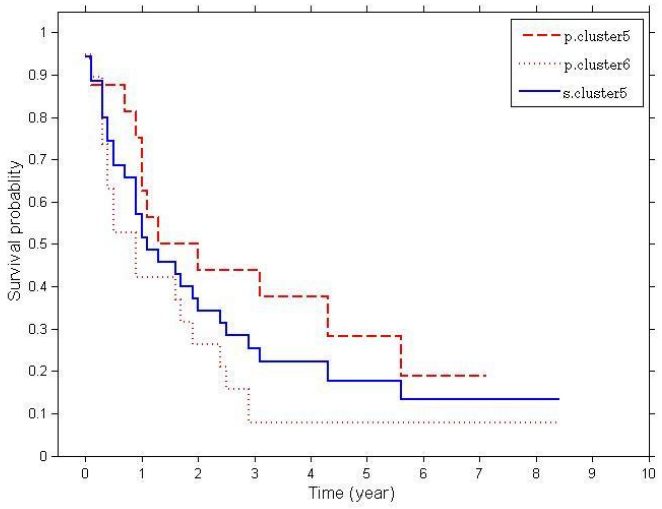

Figure 33: Kaplan-Meier survival curves correspond to the s.cluster5, p.cluster5-6 in Figure 32.

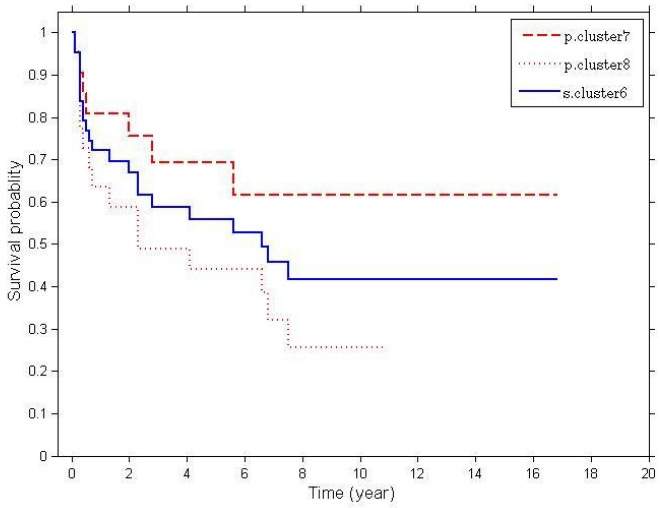

Figure 34: Kaplan-Meier survival curves correspond to the s.cluster6, p.cluster7-8 in Figure 32.

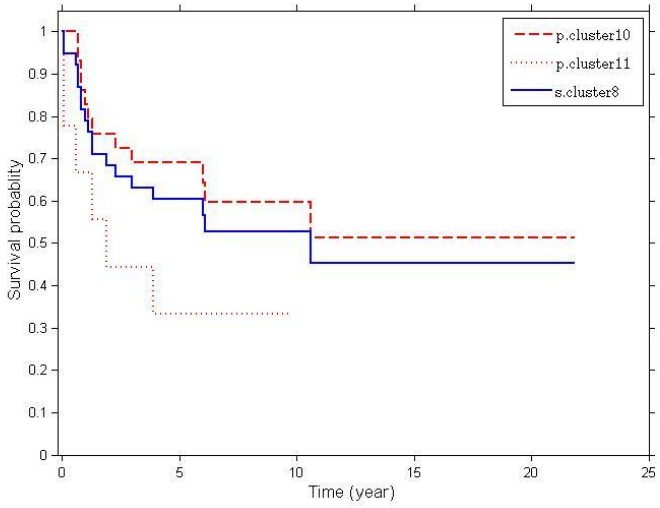

Figure 35: Kaplan-Meier survival curves correspond to the s.cluster8, p.cluster10-11 in Figure 32.

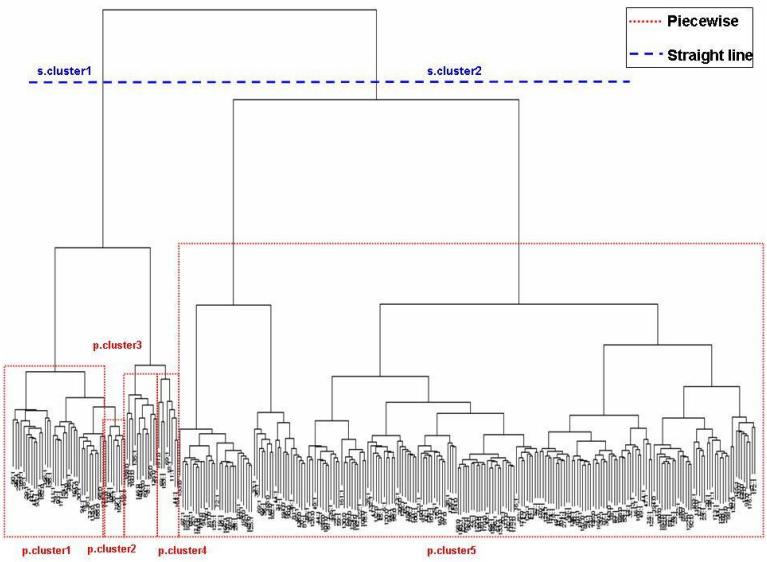

Figure 36: The HC tree corresponds to the Prostate cancer data set [Sboner et al., 2010], and the optimal cuts induced by the two approaches.

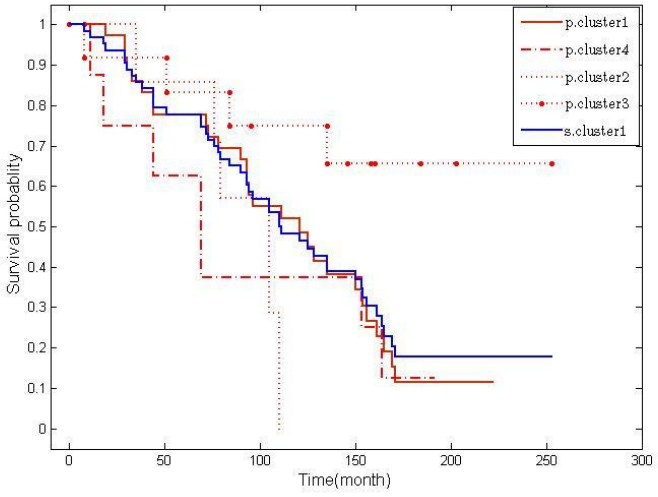

Figure 37: Kaplan-Meier survival curves correspond to the s.cluster1, p.cluster1-4 in Figure 36.

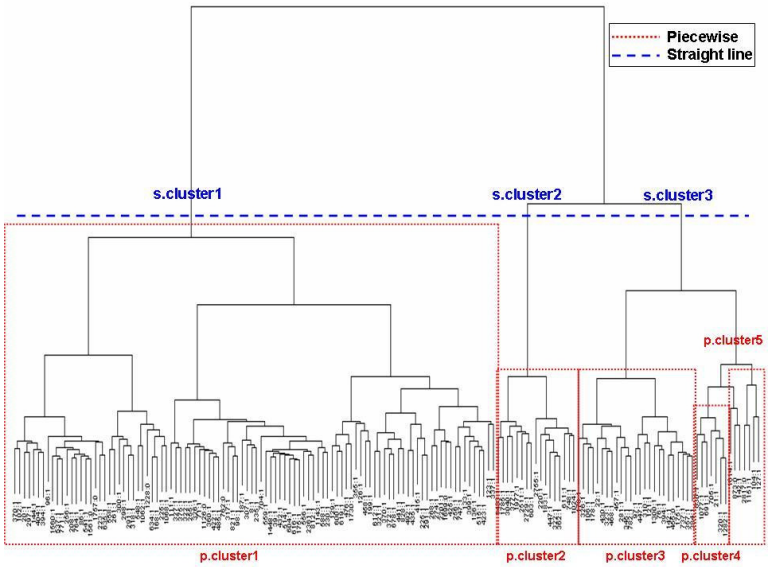

Figure 38: The HC tree corresponds the GBM data set [Verhaak et al., 2010], and the optimal cuts induced by the two approaches.

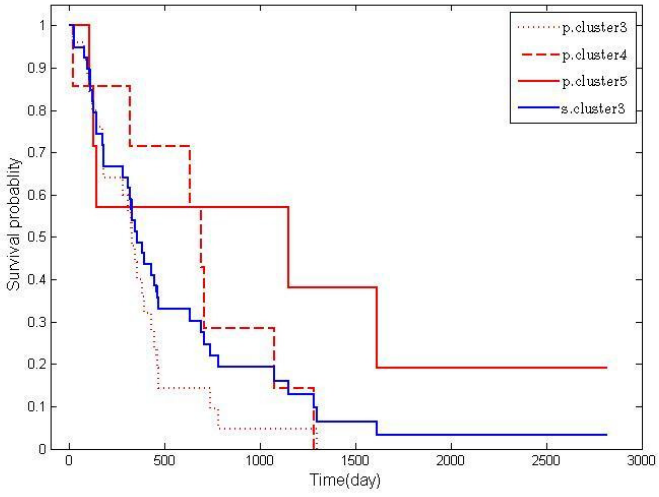

Figure 39: Kaplan-Meier survival curves correspond to the s.cluster2, p.cluster3-5 in Figure 38.

Appendix C

This appendix provides supplementary information not mentioned in Chapter 6.

Array CGH data preprocessing

After computing log2 ratios, missing values were imputed using a k-nearest neighbour algorithm implemented in the R-package impute available from Bioconductor. Missing values were imputed if values of a particular feature were available from more than 30% of all experiments. By applying this imputation procedure, the total number of features was reduced to 173,367 features. Afterwards, CGH profiles were wave bias corrected by regressing them on a calibration set containing 16 normal profiles to improve detection of aberrations [van de Wiel et al., 2009b]. In a last preprocessing step, microarray data was global median normalized and tumor % corrected using an approach described by van de Wiel et al. [2005]. Subsequently, copy number profiles were inspected visually. The median cellularity of remaining 75 samples is 60%. The final data matrix that has been used for downstream analysis was of size $X \in R^{173367 \times 75}$. Normalization, cellularity correction and segmentation were performed with the R-package CGHcall was used for preprocessing and segmentation.

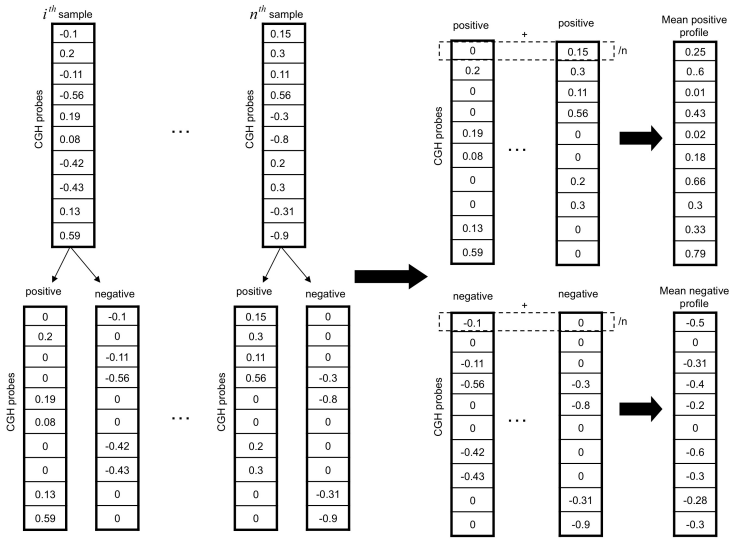

Figure 40: Schemata of the averaged CGH profile generation from a given segmented data. For each sample its segmented data values were first divided into positive and negative parts, roughly equal to gain and loss. The positive values across samples were averaged to calculate mean positive segmented values. The mean negatives segmented values were generated in similar ways. Finally, they were plotted together in barplots to illustrate the averaged DNA copy number aberration patterns in a given segmented data.

The DNA copy number entropy vs. Tumour cells/areas

To accompany the visualization in Figure 42, we also conducted a formal statistical test. Namely, we tested the statistical significance of the group differences in terms of entropy in the presence of the tumor cells/areas by a simple regression. The DNA copy number entropy used as response variable, whereas the treatment-arm indicator and the tumor cells/areas were used as independent covariates. Regression result showed that the treatment-arm indicator was the strong predictor (coeff. = 0.216, p = 0.004), while no significant association observed for the tumor cells/areas (coeff. = -0.002, p = 0.29). Thus, we confirmed that the observed differences between the two treatment arms indeed were not due to the tumor cells/areas.

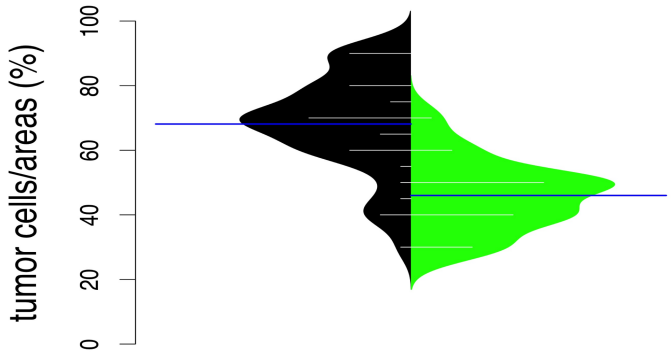

Figure 41: Bean plots comparing the tumor cells/areas distributions between the treatment groups. The shape and the mean (blue line) of tumor cells/areas are relatively different between the treatment groups. S (right) appears to be composed of samples with low tumor cells/areas, and opposite holds for CS (left).

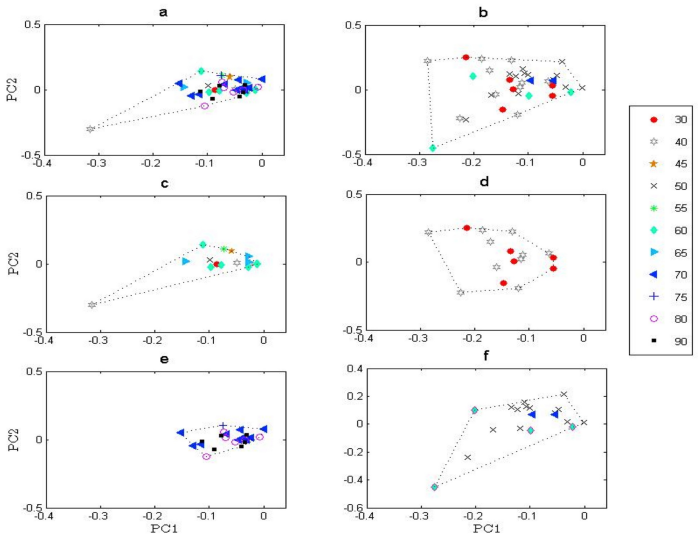

Figure 42: The tumor cells/areas difference vs. the DNA copy number entropy difference. Top row show the distributions of samples from CS (a) and S (b) in the space defined by their first two principal components obtained from the segmented data matrix. We also divided samples in each treatment arm into two groups (low and high cellularity) according to the tumor cells/areas, the median used as cutoff. Then, samples were again projected onto the space spanned by the first two principal components. We observed that there was no considerably difference in scatterness between the high and the low tumor cells/areas groups within each treatment arm, eg. (c) vs (e) and (d) vs (f). Also, no considerably difference in scatterness between the low tumor cells/areas group (c) in CS and the high tumor cells/areas group (f) in S was observed. Hence, we concluded that the observed difference between the treatment groups were not due to the artifact of the difference in the tumor cells/areas. The differences were indeed caused by the treatment group specific aberration patterns in the CGH profiles. In each panel a point denotes a sample, and the shape and color corresponds to one of the tumor cell/areas percentages shown in the legend.

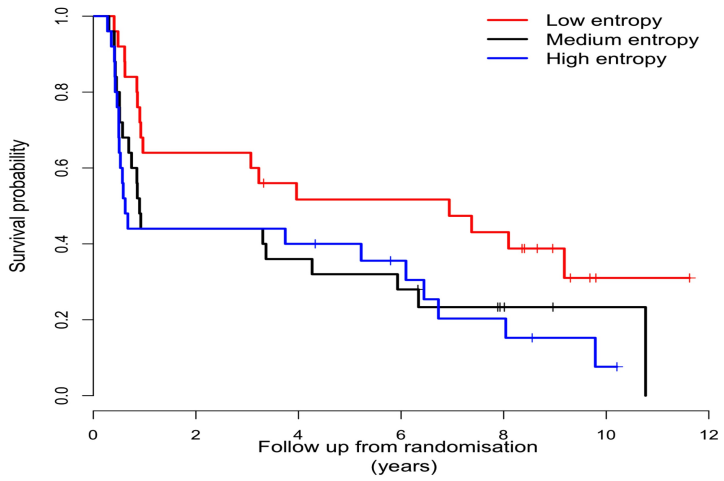

Figure 43: Association of DNA copy number entropy with cancer specific survival. The whole cohort was divided into three equal sized patient groups based on the DNA copy number entropy values. Observe that the low entropy group (n=25) corresponds to better prognosis (median (range) survival time: 3.96 (0.41-11.62) years), largely overlapped the medium (n = 25, 0.90 (0.30-10.77) years) and high entropy (n = 25, 0.62 (0.28-10.21) years) groups, however, exhibit poor prognoses.

References

A.A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Jr. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Bostein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000. 10, 54, 56

W.H. Allum, S.P. Stenning, J. Bancewicz, P.I. Clark, and R.E. Langley. Long-term results of a randomised trial of surgery with or without preoperative chemotherapy in oesophageal cancer. *J Clin Oncol*, 27:5062–5067, 2009. 86, 87, 95

V. Almendro, Y. K. Cheng, A. Randles, S. Itzkovitz, A. Marusyk, E. Ametller, X. Gonzalez-Farre, M. Mu noz, H. G. Russnes, A. Helland, I. H. Rye, A. L. Borresen-Dale, R. Maruyama, A. van Oudenaarden, M. Dowsett, R. L. Jones, J. Reis-Filho, P. Gascon, M. Gönen, F. Michor, and K. Polyak. Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell reports*, 6:514–527, 2014. 86, 96

H. Armendariz, M.A. Barbieri abd D. Freigeiro, F. Lastiri, M.S. Felice, and E. Dibar. Treatment strategy and long-term results in pediatric patients treated in two consecutive AML-GATLA trials. *Leukemia*, 19:2139–2142, 2005. 8

E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLosS Biol.*, 2:511–522, 2004. 9, 54, 55, 68, 70

A. E. Baya and P. M. Granitto. Clustering gene expression data with a penalized graph-based metric. *BMC Bioinformatics*, 12:2, 2011. 10, 56, 110

D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature*, 8:816–824, 2002. vii, viii, xi, 11, 54, 55, 110, 113

A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Beuno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Suqarbeker, and M. Meyserson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, 98:13790–13795, 2001. xi, 10, 54, 56, 114

M. Blink, T. D. Buitenkamp, M. M. van den Heuvel-Eibrink, A. A. Danen van Oorschot, V. de Haas, D. Reinhardt, J. H. Klusmann, M. Zimmermann, M. Devidas, A. J. Carroll, G. Basso, A. Pession, H. Hasle, R. Pieters, K. R. Rabin, S. Izraeli, and C.M. Zwaan. Frequency and prognostic implications of JAK 1-3 aberrations

in down syndrome acute lymphoblastic and myeloid leukemia. *Leukemia*, 25: 1365–1368, 2011. 8

N. Bolshakova, D. R. Haynor, and W. L. Ruzzo. Cluster validation for gene expression data. *Bioinformatics*, 19:2494–2495, 2003. 70

A. L. Boulesteix. Maclinical: Class prediction based on microarray data and clinical parameters. *Bioconductor*, 2012. 39, 47

A.-L. Boulesteix and W. Sauerbrei. Added predictive value of high-throughput molecular data to clinical data and its validation. *Brief Bioinform.*, 12:215–229, 2011. 6, 21

A.-L. Boulesteix and C. Strobl. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Med Res Methodol.*, 9:85, 2009. 31

A.-L. Boulesteix, C. Porzelius, and M. Daumer. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24:1698–1706, 2008. 5, 6, 8, 21, 27, 28, 29, 30, 32, 34, 35, 100

M. H. Bovelstad, S. Nygard, and O. Borgan. Survival prediction from clinico-genomic models - a comparative study. *BMC Bioinformatics*, 10:413, 2009. 6, 21

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 23, 24, 40

T. E. Buffart, D. Israeli, M. Tijssen, S. J. Vosse, A. Mrsić, G. A. Meijer, and B. Ylstra. Across array comparative genomic hybridization: a strategy to reduce reference channel hybridizations. *Genes Chromosomes Cancer*, 47:994–1004, 2008. 88

L. Bullinger, K. Dohner, E. Bair, S. Fröhling, R. F. Schlenk, R. Tibshirani, H. Döhner, and J. R. Pollack. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med.*, 350:1605–1616, 2004. 112

Jr. W. E. Burak. Is neoadjuvant therapy the answer to adenocarcinoma of the esophagus? *Am J Surg*, 186:296–300, 2003. 13, 86

R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501:338–345, 2013. 1, 2, 4, 86

K.-A. Lê Cao, E. Meugnier, and G. J. McLachlan. Integrative mixture of experts to combined clinical factors and gene markers. *Bioinformatics*, 24:1698–1706, 2008. 5, 8, 21, 27, 28, 29, 35

Z. Y. Chen, W. Z. Zhong, X. C. Zhang, J. Su, X. N. Yang, Z. H. Chen, J. J. Yang, Q. Zhou, H. H. Yan, S. J. An, H. J. Chen, B. Y. Jiang, T. S. Mok, and Y. L. Wu. EGFR mutation heterogeneity and the mixed response to EGFR tyrosine kinase inhibitors of lung adenocarcinomas. *Oncologist*, 17:978–985, 2012. 86

E. A. Coenen, S. C. Raimondi, J. Harbott, M. Zimmermann, T. A. Alonzo, A. Auvrignon, H. B. Beverloo, M. Chang, U. Creutzig, M. N. Dworzak, E. Forestier, B. Gibson, H. Hasle, C. J. Harrison, N. A. Heerema, G. J. Kaspers, A. Leszl, N. Litvinko, L. Lo Nigro, A. Morimoto, C. Perot, D. Reinhardt, J. E. Rubnitz, F. O. Smith, J. Stary, I. Stasevich, S. Strehl, T. Taga, D. Tomizawa, D. Webb, Z. Zemanova, R. Pieters, C. M. Zwaan, and M.M. van den Heuvel-Eibrink. Prognostic significance of additional cytogenetic aberrations in 733 de novo pediatric 11q23/MLL-rearranged AML patients: results of an international study. *Blood*, 117:7102–7111, 2011. 8

S. L. Cooke, C. K. Ng, N. Melnyk, M. J. Garcia, T. Hardcastle, J. Temple, S. Langdon, D. Huntsman, and J. D. Brenton. Genomic analysis of genetic heterogeneity and evolution in high-grade serous ovarian carcinoma. *Oncogene*, 29:4905–4913, 2010. 86

U. Creutzig, M. Zimmermann, J. Ritter, D. Reinhardt, J. Hermann, G. Henze, H. Jurgens, H. Kabisch, A. Reiter, H. Riehm, H. Gadner, and G. Schellong. Treatment strategies and long-term results in paediatric patients treated in four consecutive aml-bfm trials. *Leukemia*, 19:2130–2142, 2005. 8

K. D. Crew and A. I. Neugut. Epidemiology of upper gastrointestinal malignancies. *Semin. Oncol*, 31:450–464, 2004. 12, 86

D. Cunningham, W. H. Allum, S. P. Stenning, J. N. Thompson, C. J. van de Velde, M. Nicolson, J. H. Scarffe, F. J. Lofts, S. J. Falk, T. J. Iveson, D. B. Smith, R. E. Langley, M. Verma, S. Weeden, Y. J. Chua, and MAGIC Trial Participants. Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. *N. Engl. J. Med*, 355:11–20, 2006. 86

L. Ding, T. J. Ley, D. E. Larson, C. A. Miller, D. C. Koboldt, J. S. Welch, J. K. Ritchey, M. A. Young, T. Lamprecht, M. D. McLellan, J. F. McMichael, J. W. Wallis, C. Lu, D. Shen, C. C. Harris, D. J. Dooling, R. S. Fulton, L. L. Fulton, K Chen, H. Schmidt, J. Kalicki-Veizer, V. J. Magrini, L. Cook, S. D. McGrath, T. L. Vickery, M. C. Wendl, S. Heath, M. A. Watson, D. C. Link, M. H. Tomasson, W. D. Shannon, J. E. Payton, S. Kulkarni, P Westervelt, M. J. Walter, T. A. Graubert, E. R. Mardis, R. K. Wilson, and J. F. DiPersio. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481:506–510, 2012. 86

D. Dotan-Cohen, A. A. Melkman, and S. Kasif. Hierarchical tree snipping: clustering guided by prior knowledge. *Bioinformatics*, 23:3335–3342, 2007. 9, 55, 109, 110

A. M. Dulak, S. E. Schumacher, J. van Lieshout, Y. Imamura, C. Fox, B. Shim, A. H. Ramos, G. Saksena, S. C. Baca, J. Baselga, J. Tabernero, J. Barretina, P. C. Enzinger, G. Corso, F. Roviello, L. Lin, S. Bandla, J. D. Luketich, A. Pennathur, M. Meyerson, S. Ogino, R. A. Shivdasani, D. G. Beer, T. E. Godfrey, R. Beroukhim, and A. J. Bass. Gastrointestinal adenocarcinomas of the esophagus, stomach, and colon exhibit distinct patterns of genome instability and oncogenesis. *Cancer Res*, 72:4383–4393, 2012. viii, 88, 89, 90, 91, 95

A. M. Dulak, P. Stojanov, S. Peng, M. S. Lawrence, C. Fox, C. Stewart, S. Bandla, Y. Imamura, S. E. Schumacher, E. Shefler, A. McKenna, S. L. Carter, K. Cibulskis, A. Sivachenko, G. Saksena, D. Voet, A. H. Ramos, D. Auclair, K. Thompson, C. Sougnez, R. C. Onofrio, C. Guiducci, R. Beroukhim amd Z. Zhou, L. Lin, J. Lin, R. Reddy, A. Chang, R. Landrenau, A. Pennathur, S. Ogino, J. D. Luketich, T. R. Golub, S. B. Gabriel, E. S. Lander, D. G. Beer, T. E. Godfrey, G. Getz, and A. J. Bass. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet*, 45:478–486, 2013. 94

D. Dunkler, S. Michiels, and M. Schemper. Gene expression profiling: Does it add predictive accuracy to clinical characteristics in cancer prognosis? *Eur J Cancer*, 12:153–157, 2007a. 6, 7, 21

D. Dunkler, S. Michiels, and M. Schemper. Gene expression profiling: Does it add predictive accuracy to clinical characteristics in cancer prognosis? *Eur J Cancer*, 12:745–751, 2007b. 35

L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *PNAS*, 103:5923–5928, 2006. 5

A. Frankel, N. Armour, D. Nancarrow, L. Krause, N. Hayward, G. Lampe, B. M. Smithers, and A. Barbour. Genome-wide analysis of esophageal adenocarcinoma yields specific copy number aberrations that correlate with prognosis. *Genes Chromosomes Cancer*, 53:324–338, 2014. 94, 95

M. E. Futschik, M. Sullivan, A. Reeve, and N. Kasabov. Prediction of clinical behaviour and treatment for cancers. *Appl Bioinformatics*, 2:53–58, 2003. 20

T. A. Gerds and M. A. van de Wiel. Confidence scores for prediction models. *Biom J*, 53:259–274, 2011. 49

J. Goeman, R. Meijer, and N. Chaturverdi. L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. *R package*, 2012. 39

J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwellingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20:93–99, 2004. 11, 60, 64, 80

X. Y. Goh, J. R. Rees, A. L. Paterson, S. F. Chin, J. C. Marioni, V. Save, M. O'Donovan, P. P. Eijk, D. Alderson, B. Ylstra, C. Caldas, and R. C. Fitzgerald. Integrative analysis of array-comparative genomic hybridisation and matched gene expression profiling data reveals novel genes with prognostic significance in oesophageal adenocarcinoma. *Gut*, 60:1317–1326, 2011. viii, 88, 89, 90, 91, 94, 95

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999. 5

L. Goodman and W. Kruskal. Measures of associations for cross-validations. *J. Am. Stat. Assoc*, 49:732–764, 1954. 70

F. E. Harrel, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *JAMA*, 247:2543–2546, 1982. 59, 73

T. Hastie, R. Tibshirani, B. Narasimhan, and G. Chu. pamr: Prediction analysis for microarrays. *Bioconductor*, 2011. 39

C. Hennig. Flexible procedures for clustering. *R package*, 2010. 70, 110

Z. Huang. Clustering large data sets with mixed numeric and categorical values. *The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 16–27, 1997. 22, 23

L. Hubert and J. Schultz. Quadratic assignment as a general data-analysis strategy. *Br J Math Statist Psych*, 29:190–241, 1976. 70

L. Huiqing. *Effective use of data mining technologies on biological and clinical data*. PhD thesis, National University of Singapor, School of computing, 2004. 1

T. Ideker, J. Dutknowski, and L. Hood. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, 144:860–863, 2011. 55

A. V. Ivshina, J. George, O. Senko, B. Mow, T. C. Putti, J. Smeds, T. Lindahl, Y. Pawitan, P. Hall, H. Nordgren, J. E. Wong, E. T. Liu, J. Bergh, V. A. Kuznetsov, and L. D. Miller. Genetic reclassification of histologic gade delineates new clinical subtypes of breast cancer. *Cancer Res*, 66:10292–10301, 2006. 45

M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer, and A.-L. Boulesteix. Over-optimism in bioinformatics: an illustration. *Bioinformatics*, 16:1990–1998, 2010. 5, 31

JOUR. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609–615, 2011. 78

V. A. Kapp and R. Tibshirani. Are clusters found in one dataset present in another dataset? *Biostatistics*, 8:9–31, 2007. 65, 71

C. A. Klein and N. H. Stoecklein. Lessons from an aggressive cancer: evolutionary dynamics in esophageal carcinoma. *Cancer Res*, 69:5285–5288, 2009. 13, 86

D. Krag, D. Weaver, and T. Ashikaga. The sentinel node in breast cancer a multicenter validation study. *N Engl J Med.*, 339:941–946, 1998. 20

O. Krijgsman, D. Israeli, J. C. Haan, H. F. van Essen, S. J. Smeets, P. P. Eijk, R. D. Steenbergen, K. Kok, S. Tejpar, G. A. Meijer, and B. Ylstra. CGH arrays compared for DNA isolated from formalin-fixed, paraffin-embedded material. *Genes Chromosomes Cancer*, 51:344–352, 2012. 88

O. Krijgsman, D. Israeli, H. F. van Essen, P. P. Eijk, M. L. Berens, C. H. Mellink, A. W. Nieuwint, M. M. Weiss, R. D. Steenbergen, G. A. Meijer, and B. Ylstra. Detection limits of DNA copy number alterations in heterogeneous cell populations. *Cell Oncol*, 36:27–36, 2013. 88

R. Kustra and A. Zagdanski. Incorporating gene ontology in clustering gene expression data. *Proceeding CBMS '06 Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*, pages 555–563, 2006. 55

P. Langfelder, B. Zhang, and S. Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24:719–720, 2008. 54, 55

J. T. Leek. tspair: Top scoring pairs for microarray classification. *Bioconductor*, 2009. 39

H. Liang and H. G. Zou. Improved AIC selection strategy for survival analysis. *Comput Stat Anal.*, 52:2538–2548, 2008. 58, 70

C. C. Maley, P. C. Galipeau, J. C. Finley, V. J. Wongsurawat, X. Li, C. A. Sanchez, T. G. Paulson, P. L. Blount, R. A. Risques, P. S. Rabinovitch, and B. J. Reid. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet*, 38:468–473, 2006. 13, 86

A. M. Mandard, F. Dalibard, J. C. Mandard, J. Marnay, M. Henry-Amar, J. F. Petiot, A. Roussel, J. H. Jacob, P. Segol, and G. Samama. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. clinicopathologic correlations. *Cancer*, 73:2680–2688, 1994. 13, 87

MRC Medical Research Council Oesophageal Cancer Working Group. Surgical resection with or without preoperative chemotherapy in oesophageal cancer: a randomised controlled trial. *Lancet*, 359:1727–1733, 2002. 2, 4, 13

L. M. Merlo, J. W. Pepper, B. J. Reid, and C. C. Maley. Cancer as an evolutionary and ecological process. *Nat Rev Cancer*, 6:924–935, 2006. 13, 86

L. M. Merlo, N. A. Shah, X. Li, P. L. Blount, T. L. Vaughan, B. J. Reid, and C. C. Maley. A comprehensive survey of clonal diversity measures in barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer Prev Res (Phila)*, 3:1388–1397, 2010. 13, 86

D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. e1071: Misc functions of the ddepartment of statistics (e1071), TU Wien. *Bioconductor*, 2012. 39

S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 4958:488–492, 2005. 5, 34

G. W. Milligan. A Monte-Carlo study of 30 internal criterion measures for cluster-analysis. *Psychometrica*, 46(2):187–195, 1981. 58

E. A. Mroz, A. D. Tward, C. R. Pickering, J. N. Myers, R. L. Ferris, and J. W. Rocco. High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer*, 119:3034–3042, 2013. 2, 86, 95, 97

G. G. Mullighan, L. A. Phillips, X. Su, J. Ma, C. B. Miller, S. A. Shurtleff, and J. R. Downing. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*, 322:1377–1380, 2008. 86

S. Navlakha, J. White, N. Nagarajan, M. Pop, and C. Kingsford. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. *J Comp Biol.*, 17:503–516, 2010. 9, 56

R. J. Nevins, S. E. Huang, and H. Dressman. Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Hu Mol Genet.*, 43:745–751, 2003. 6, 20

A. Obulkasim, G. A. Meijer, and M. A. van de Wiel. Stepwise classification of cancer samples using clinical and molecular data. *BMC Bioinformatics*, 12:422, 2011. 6, 34, 39, 41, 58, 59, 72

A. Obulkasim, G. A. Meijer, and M. A. van de Wiel. Semi-supervised adaptive-height snipping of the hierarchical clustering tree. *BMC Bioinformatics*, 16:15, 2015. 68

M. Y. Park and T. Hastie. glmpath: L1 regularization path for generalized linear models and cox proportional hazards model. *Bioconductor*, 2013. 39

G. Pasello, S. Agata, L. Bonaldi, A. Corradin, M. Montagna, R. Zamarchi, A. Parenti, M. Cagol, G. Zaninotto, A. Ruol, E. Ancona, A. Amadori, and D. Saggioro. DNA copy number alterations correlate with survival of esophageal adenocarcinoma patients. *Mod Pathol*, 22:58–65, 2009. 94, 95

T. G. Paulson, C. C. Maley, X. Li, H. Li, C. A. Sanchez, D. L. Chao, R. D. Odze, T. L. Vaughan, P. L. Blount, and B. J. Reid. Chromosomal instability and copy number alterations in barrett's esophagus and esophageal adenocarcinoma. *Cancer Res*, 22:58–65, 2009. 94

H. Pohl and H. G. Welch. The role of overdiagnosis and reclassification in the marked increase of esophageal adenocarcinoma incidence. *J Natl Cancer Inst*, 97:142–146, 2005. 13, 86

S. L. Pomeroy, P. Tamayo, and M. Gaasenbeek. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002. 27, 43

N. E. Potter, L. Ermini, E. Papaemmanuil, G. Cazzaniga, G. Vijayaraghavan, I. Titley, A. Ford, P. Campbell, L. Kearney, and M. Greaves. Single cell mutational profiling and clonal phylogeny in cancer. *Genome Res*, 23:2115–2125, 2013. 1

Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple sources. *Pacific Symposium on Biocomputing*, pages 531–542, 2005. 23, 40

L. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, E. M. Hurt, H. Zhao, L. Averett, W. H. Wilson L. Yang, E. S. Jaffe, R. Siomon, R. D. Klausner, J. Powell, P. L. Duffey, D. L. Longo, T. C. Greiner, D. D. Weisenburger, W. G. Sanger D. J. Dave, J. C. Lynch, J. Vose, J. O. Armitage, E. Montserrat, A. López-Guillermo, T. M. Grogan, T. P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, and L. M. Staudt. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *N Engl J Med.*, 346:1937–1947, 2002. xi, 116

P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med.*, 25:127–141, 2006. 10, 56

N. A. Saunders, F. Simpson, E. W. Thompson, M. M. Hill, L. Endo-Munoz, G. Leggatt, R. F. Minchin, and A. Guminski. Role of intratumoural heterogeneity in cancer drug resistance: molecular and clinical perspectives. *EMBO Mol Med*, 4: 675–684, 2012. 2, 86

A. Sboner, F. Demichelis, S. Calza, Y. Pawitan, S. R. Setlur, Y. Hoshida, S. Perner, H.-O. Adami, K. Fall, L. A. Mucci, P. W. kantoff, M. Stamfer, S.-O. Andersson, E. Varenhorst, J.-E. Johansson, M. B. Gerstein, T. R. Golub, M. A. Rubin, and O. Andrén. Molecular sampling of prostate cancer: a dilemma for predicting disease progression. *BMC Med Genomics*, 3:3–8, 2010. xi, 54, 57, 68, 118

M. Schumacher, N. Holländer, G. Schwarzer, H. Binder, and W. Sauerbrei. Prognostic factor studies. *Handbook of Statistics in Clinical Oncology*, 2 ed.:289–333, 2006. 34

S. J. Smeets, U. Harjes, W. N. van Wieringen, D. Sie, R. H. Brakenhoff, G. A. Meijer, and B. Ylstra. To dna or not to dna? that is the question, when it comes to molecular subtyping for the clinic! *Clinical Cancer Res*, 17:4959–4964, 2011. 4

R.L. Somorjai, B. Dolenko, and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, 19:1484–1491, 2003. 5

C. Soneson, H. Lilljebjörn, T. Fioretos, and M. Fontes. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics*, 11:191, 2010. 5

T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, G. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A.-L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98:10869–10874, 2001. vii, xi, 9, 10, 54, 68, 109

M.C. De Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9: 497, 2008. 9, 54, 79

J. A. Stephenson, A. Smit, and W. M. Katta. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer*, 104:290–298, 2005. 6, 21, 27, 100

A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, and D. Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21: 3896–3904, 2005. 28

W. Tang, J. Duan, J.G. Zhang, and Y.P. Wang. Subtyping glioblastoma by combining miRNA and mRNA expression data using compressed sensing-based approach. *EURASIP J Bioinform Syst Biol*, 2013:2, 2013. 5

J.N. Weinstein The Cancer Genome Atlas Research Network, E.A. Collisson, B.G. Mills, K.R.M. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J.M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, 45:1113–1120, 2013. 2

J. R. Tibshirani and B. Efron. Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol*, 1, 2002. 30

M. J. van de Vijver, Y. D. He, and L. J. van't Veer. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.*, 347:1999–2009, 2002. vii, 7, 8, 27

M. A. van de Wiel, K. I. Kim, S. J. Vosse, W. N. van Wieringen, S. M. Wilting, and B. Ylstra. CGHcall: an algorithm for calling aberrations for multiple array CGH tumor profiles. *Bioinformatics*, 23:1–7, 2005. 88, 121

M. A. van de Wiel, J. Berkhof, and W. N. van Wieringen. Testing the prediction error difference between 2 predictors. *Biostatistics*, 10:550–560, 2009a. 60

M. A. van de Wiel, R. Brosens, P. H. Eilers, C. Kumps, G. A. Meijer, B. Menten, E. Sistermans, F. Speleman, M. E. Timmerman, and B. Ylstra B. Smoothing waves in array CGH tumor profiles. *Bioinformatics*, 25:1099–1104, 2009b. 121

M. A. van de Wiel, F. Picard, W. N. Wieringen, and B. Ylstra. Preprocessing and downstream analysis and microarray DNA copy number profiles. *Brief Bioinform*, 12:10–21, 2011. 3

H. F. van Essen and B. Ylstra. High-resolution copy number profiling by array CGH using DNA isolated from formalin-fixed, paraffin-embedded tissues. *Methods Mol Biol*, 838:329–341, 2012. 88

P. van Hagen, M. C. Hulshof, J. J. van Lanschot, E. W. Steyerberg, M. I. van Berge Henegouwen, B. P. Wijnhoven, D. J. Richel, G. A. Nieuwenhuijzen, G. A. Hospers, J. J. Bonenkamp, M. A. Cuesta, R. J. Blaisse, O. R. Busch, F. J. ten Kate, G. J. Creemers, C. J. Punt, J. T. Plukker, H. M. Verheul, E. J. Spillenaar Bilgen, H. van Dekken, M. J. van der Sangen, T. Rozema, K. Biermann, J. C. Beukema, A. H. Piet, C. M. van Rij, J. G. Reinders, H. W. Tilanus, A. van der Gaast, and CROSS Group. Preoperative chemoradiotherapy for esophageal or junctional cancer. *N. Engl. J. Med*, 366:2074–2084, 2012. 86

W. N. van Wieringen and A. W. van der Vaart. Statistical analysis of the cancer cells molecular entropy using high-throughput data. *Bioinformatics*, 27:556–563, 2010. 4, 13, 61, 77, 86, 89

W. N. van Wieringen, M. A. van de Wiel, and B. Ylstra. Weighted clustering of array CGH data. *Biostatistics*, 9:484–500, 2008. 57, 77, 81

R. G. Verhaak, , K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. R. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodqson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman,

R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, and D. N. Hayes. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17:98–110, 2010. xi, 47, 76, 81, 119

C. T. Volinsky and E. A. Raftery. Baysian information criteria for censored survival models. *Biometrics*, 56:256–262, 2000. 58, 70

J. H. Ward. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc*, 58:236–244, 1963. 54

M.M. Weiss, M.A.J.A. Hermsen, G.A. Meijer, N.C.T. van Grieken, J.P.A. Baak, and E.J. Kuipers. Comparative genomic hybridisation. *Mol Pathol*, 52:243–251, 1999. 3

W. N. Wieringen. sigaR: Statistics for integrative genomics analyses in R. *Bioconductor*, 2012. 88

D.H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8:1341–1390, 1996. 48

Z. Yong, Y. Yupu, and Z. Liang. Pseudo nearest neighbor rule for pattern classification. *Expert Syst. Appl*, 36:3587–3595, 2009. 25

T. Yu, J. Li, and S. Ma. Adjusting confounders in ranking biomarkers: a model-based ROC approach. *Brief Bioinform*, 13:513–523, 2011. 81

J. Zacherl, A. Sendler, H. J. Stein, K. Ott, M. Feith, R. Jakesz, J. R. Siewert, and U. Fink. Current status of neoadjuvant therapy for adenocarcinoma of the distal esophagus. *World J Surg*, 27:1067–74, 2003. 13, 86

Summary

The coming century is surely the century of *data*. Rapid advances in biotechnology and life science made it possible to collect and process genomics data of all kinds, on scales unimaginable a few decades ago. Different molecular data profiles provide a different, partly independent and complementary molecular fingerprint of a tissue. The challenge, however, is no longer how to generate the genomics data, but rather how to analyze these. This dissertation aims at contributing to solve such analysis problems for data in the context of cancer genomics.

Two important topics in the field of bioinformatics, data integration and characterization of cancer heterogeneity, are the focus of this dissertation. It is composed of three parts. In the first two parts, we propose a general framework for data integration for the purpose of cancer genotype classification and clustering. The motivation to write these two parts stems from the fact that the biological insight harvested from either one of the individual data types is often limited. Our goal is to address the spurring need for integration of diverse data measured on the same individuals. In the third part of this thesis we discuss the potential of exploring cancer heterogeneity in clinical applications.

Chapter 1 aims at discussing some general background on the topics discussed in the thesis. This chapter serves mean of helping the reader approach the remaining chapters (Chapters 2 and 3) with more ease. In Chapter 2 existing integrative classification methods are studied and a novel integrative classifier that is designed to combine commonly available patient clinical risk factors with genomics data is introduced. The proposed approach, called stepwise classifier, addresses the major shortcoming of existing approaches: the requirement to measure genomics data for all patients. This may make the setting costly, impractical or inefficient. The stepwise classifier, however, requires genomics data to be available for a subset of patients only, namely for those that are predicted to benefit the most by measuring their molecular profiles. From the research application perspective, this approach is appealing. In Chapter 3 the R-BioConductor package `stepwiseCM`, which implements the cost efficient stepwise classification strategy introduced in Chapter 2, is presented. This package, with many easy to use functions and an elaborated vignette, can be a useful tool in research applications where, next to accuracy, efficiency is an important goal.

Part II is composed of Chapter 4 and 5. In Chapter 4 we introduce a data driven, rather than heuristic, and semi-supervised rather than fully unsupervised, cluster extraction framework from a hierarchical clustering (HC) tree. This method integrates genomics data with background information to tease out a meaningful clustering from the HC tree. No restriction is placed on the type of background information that can be used. This can be (partial) labels of samples, another type of genomics data, or patient outcome data, such as survival. The ability to accommodate the patient outcome data without discretization makes our approach stand out among its competitors. Furthermore, we extended the application of the proposed approach to optimal treatment assignment. We show that our method significantly improves treatment assignment compared to the original assignment. Hence, this approach fits very well within the individualized medicine paradigm.

The R-BioConductor package `HCsnip` and the corresponding manual form the content of Chapter 5. This package implements the semi-supervised HC tree snipping framework presented in Chapter 4.

Chapter 6 introduces a clinical study in which a subset of patients with oesophageal adenocarcinoma (OeC) recruited into the esophageal cancer trial (OEO2) is investigated. The OEO2 trial is a randomized trial in which one of the two following treatments was randomly allocated: surgery alone or, prior to surgery, two cycles of a combination cisplatin and fluorouacil. This trial was designed to evaluate whether preoperative chemotherapy followed by surgery (CS) improves survival compared to surgery alone (S) and to determine molecular differences between the two groups. We first applied the semi-supervised clustering method presented in Chapter 4 to unravel clusters that are associated with patient clinical outcome. Then, intratumor genomic heterogeneity in this study was characterized In particular, we studied whether the intratumor heterogeneity, as measured by genomic entropy, is different before and after cytotoxic chemotherapy, and is associated with OeC patient survival. We found that (i) the DNA copy number entropy is not simply a surrogate of some other pathological variable; (ii) the two patient groups (CS versus S) have differential genomic intratumoral heterogeneity; (iii) the between-tumor heterogeneity is smaller in the CS group compared that of the S group. We conclude that because of (tentative) molecular effect of chemotherapy OeCs after chemotherapy tend to have DNA copy number profiles in which the aberrations were found more frequently at relatively similar locations making them more homogenous as measured by the DNA copy number entropy compared to chemo-naïve OeCs. To our knowledge, this is the first study to show that cytotoxic chemotherapy appears to effect the tumour genotype (DNA copy number) in cases where changes in the histological phenotype were not visible to naked eye.

To sum up, in this thesis we have developed general frameworks for integrative classification and clustering, where commonly available patients clinical risk factors and high-dimensional genomics data are appropriately combined. Instead of limiting ourselves to very specific topics, we took different perspectives. First, we focused on practical integrative classifiers. We developed a cost-effective (and possibly more patient-friendly) integrative classifier, which performed at least as well as alternative integrative classifiers. Second, we proposed to snip the HC tree at variable heights to extract clusters while using available patient clinical data as guidance. It is a semi-supervised approach that is able to generate meaningful clusters. Aforementioned integrative approaches are quite general. Although the integration of clinical and genomics data is the primary focus of this thesis, application to combinations of two types of genomic data, or even multiple types of data is usually feasible under our integration frameworks. Third, we proposed to quantify intratumor heterogeneity via genomic entropy, and use it to examine the association with patient survival. To our knowledge this is the first study to use genomic entropy in a clinical study. All our methods have been implemented in R-(Bioconductor)-packages, which together form our contribution to cancer bioinformatics.

Samenvatting

Deze eeuw is zeker de eeuw van de *data*. Snelle ontwikkelingen op het gebied van biotechnologie en levenswetenschappen hebben het mogelijk gemaakt op vele manieren genomischedata te verzamelen op een schaal die een paar decenia geleden niet voor mogelijk werd gehouden. Deze technieken genereren een vingerafdruk van de kankerweefsel, die deels onafhankelijk en deels complementair is. De uitdaging ligt echter niet meer in hoe genomische data te verzamelen, maar in hoe haar te analyzeren.

Dit proefschrift richt zich op twee belangrijke onderwerpen in de bioinformatica: data integratie en tumor heterogeniteit. Het bestaat uit drie gedeeltes. In de eerste twee delen presenteren we een algemeen kader voor het integreren van data voor tumorgenotype classificatie en -clustering. De motivatie voor deze twee delen is het feit dat biologische inzichten uit afzonderlijke data types (bijv. genomisch en klinisch) vaak beperkt zijn. Ons doel hier is dus het integreren van verschillend genomische data met clinische gegevens.

Hoofdstuk 1 is gericht op het bespreken van enkele algemene achtergrondinformatie over de onderwerpen besproken in het proefschrift. Dit hoofdstuk dient ook als inleiding voor de hieropvolgende hoofdstukken 2 en 3. In hoofdstuk 2 worden bestaande integratieve classificatiemethodes geanalyseerd samen met een nieuwe methode die ontworpen is om standaard risicofactoren te combineren met genomische data. De nieuwe methode, genaamd stepwise classifier, presenteert een oplossing voor een problematische vereiste van bestaande methodes, namelijk de aanwezigheid van genomische data van iedere patient. Die vereiste is vaak niet haalbaar of maakt de huidige methodes duur en inefficient. De stepwise classifier vereist echter slechts aanwezigheid van genomische data in een subset van patinten waarvan wordt voorspeld dat ze het meeste baat hebben van de moleculaire metingen. Ook vanuit een onderzoek oogpunt is dit aantrekkelijk. In hoofdstuk 3 wordt het R-BioConductor package `stepwiseCM` gepresenteerd, hetgeen deze kostenbesparende classificatiestrategie implementeert. Dit package bevat tal van gebruikersvriendelijke functies en een handleiding. Het kan een nuttige ondersteuning zijn bij onderzoek toepassingen waar efficientie, naast accuraatheid is.

Deel II bestaat uit hoofdstukken 4 en 5. In hoofdstuk 4 introduceren wij een niet-heuristische methode om clusters uit een hierarchische clustering (HC) structuur te extraheren.. Deze methode tracht betekenisvolle clusters te identificeren met behulp van integratie van genomische data met achtergrondinformatie. Deze achtergrondinformatie kan van elk type zijn, bijvoorbeeld sample labels, andersoortige genomische data of ziekteverloop (incl overlevingsdata). Vooral het vermogen om niet-gediscretiseerde ziekteverloopgegevens te gebruiken maakt deze methode uniek. Bovendien hebben we de methode uitgebreid met een in-silico optimalisatie van behandelingsvoorschrift. We laten zien dat onze methode het behandelingsvoorschrift significant kan verbeteren vergeleken met het oorspronkelijke voorschrift. Dezeaanpak past geheel in de trend van personalized medicine. Het R-BioConductor package `HCsnip` wordt beschreven in hoofdstuk 5. Dit package implementeert het gedeeltelijk gesuperviseerde HC structuur snipping platform zoals beschreven in hoofdstuk 4.

Hoofdstuk 6 beschrijft een klinische studie waarbij een subset van patiënten met slokdarm adenocarcinoma die deel uitmaken van de slokdarmkankertrial OEO2 wordt bestudeerd. Deze trial is gerandomiseerd voor é'en der volgende behandelingen: alleen maar chirurgie (C) of, vóór chirurgie, twee cycli van cisplatinum en fluoruracil (CC). Deze trial is ontwikkeld om een eventueel verschil in overleving te detecteren tussen deze twee groepen alsmede moleculaire verschillen tussen de twee groepen te determineren . Wei hebben de semi-gesuperviseerde clusteringmethode uit hoofdstuk 4 op deze patiëntengroep toegepast, om clusters te identificeren die zijn geassocieerd met ziekteverloop. Daarnaast hebben we genomische tumorheterogeniteit bestudeerd. We hebben onderzocht of tumorheterogeniteit, gemeten aan de hand van genomische entropie, verschilt voor en na chemotherapie, en geassocieerd is met overleving. We vonden dat (i) DNA copy number entropie niet een surrogaat is voor een ander ziektekenmerk; (ii) dat de tumoren uit de twee patiënengroepen van de OEO2 trial verschilen in genomische tumorheterogeniteit; (iii) De heterogeniteit tussen tumoren is kleiner in de CS groep vergeleken met de S groep. We concluderen dit vanwege de (mogelijke) moleculaire effecten van de chemotherapie, OeCs na chemotherapie hebben vaker DNA copy number profielen waarin aberraties vaker in relatief overeenkomende locaties werden gevonden, wat ze meer homogeen maakt, zoals ook gemeten door het DNA copy number entropie vergeleken met de chemo-naieve OeCs. Voor zover wij weten, is dit de eerste studie die laat zien dat cytologische chemotherapie een effect lijkt te hebben op het tumor genotype (DNA copy number) bij patiënten waarin veranderingen in het histologische fenotype niet zichtbaar waren met het blote oog.

Samenvattend, in dit proefschrift beschrijven wij een algemeen kader voor integratieve classificatie en clustering, waarbij algemeen toegankelijke klinische risicofactoren en hoogdimensionele genoomdata op juiste wijze worden gecombineerd. We hebben hierbij gebruik gemaakt van verschillende invalshoeken. In de eerste plaats gingen we uit van een practisch haalbare integratieve classifier. Ontwikkelden we een kosten-effectieve (en mogelijk meer patiëntvriendelijke) integratieve classifier, die uitgevoerd tenminste als alternatief integratieve classifiers. In de tweede plaats ontwikkelden we een manier om de HC structuur op te delen door op verschillende hoogtes te scheiden, waarbij we clinische data gebruikten als richtsnoer. Deze semi-supervised aanpak kan relevante clusters identificaren. De ontwikkelde aanpak is breed toepasbaar. Hoewel de primaire focus van dit proefschrift ligt op de integratie van clinische en genoombrede data, is onze aanpak ook toepasbaar op twee soort genomische data, of is uitbreiding naar drie of meer type data mogelijk. In de derde plaats beschrijven we een methode om intratumor heterogeniteit te meten aan de hand van genomische entropie, en gebruiken we deze metingen om de relatie met overleving van patiënten te bestuderen. Voor zover we weten is dit de eerste keer dat genomische entropie wordt gebruikt als clinische parameter. Al onze methoden gemplementeerd in R- (Bioconductor) -pakketten, die samen onze bijdrage aan kanker bioinformatica.

List of Publications

1. J. de Rooij, E. Beuling, **A. Obulkasim** et al. (2015) Recurrent deletions of IKZF1 in pediatric acute myeloid leukemia. *Haematologica*, Accepted.

2. J.-H. Klusmann, S. Emmrich, F. Engeland, M. El-Khatib, K. Henke, K., **A. Obulkasim** et al. (2015) miR-139 controls translation in myeloid leukemia through EIF4G2. *Oncogene*, Accepted.

3. J. E. Katsman-Kuipers, **A. Obulkasim**, C. M. Zwaan et al. (2015) Classification of pediatric acute myeloid leukemia based on miRNA expression. Submitted.

4. J. D. E. de Rooij, E. Beuling, M. M. van den Heuvel-Eibrink, **A. Obulkasim** et al. (2015) Recurrent deletions of IKZF1 in pediatric acute myeloid leukemia. Submitted.

5. **A. Obulkasim**, B. Ylstra, D. F. van Essen et al. (2015) Reduced genomic tumor heterogeneity after neoadjuvant chemotherapy IS related to favorable outcome in patients with oesophageal adenocarcinoma. Submitted.

6. **A. Obulkasim**, J. C. Earls, J. A. Eddy et al. (2015) Subtype prediction in pediatric acute myeloid leukemia: classification using differential network rank conservation revisited. Submitted.

7. **A. Obulkasim**, M. A. van de Wiel (2015) HCsnip: an R package for semi-supervised snipping of the hierarchical clustering tree. *Cancer Informatics*, 14: 1-19.

8. **A. Obulkasim**, G. A. Meijer, M. A. van de Wiel (2015) Semi- supervised adaptive-height snipping of the hierarchical clustering tree. *BMC Bioinformatics* 16:15.

9. M. Rezghi, **A. Obulkasim** (2014) Noise-free principal component analysis: an efficient dimension reduction technique for high dimensional molecular data. *Expert System with Applications*, 41: 7797-7804.

10. **A. Obulkasim**, M. A. van de Wiel (2014) stepwiseCM: An R package for stepwise classification of cancer samples using clinical and molecular Data. *Cancer Informatics*, 13: 1-11.

11. B. Carvalho, A. H. Sillars-Hardebol, C. Postma, S. Mongera, J. Terhaar Sive Droste, **A. Obulkasim** et al. (2012) Colorectal adenoma to carcinoma progression is accompanied by changes in gene expression associated with ageing, chromosomal instability, and fatty acid metabolism. *Cell Oncology*, 35(1): 53-63.

12. **A. Obulkasim**, M. A. van de Wiel, G. A. Meijer (2011) Stepwise classification of cancer samples using clinical and molecular data. *BMC Bioinformatics*, 12:422.

13. **A. Obulkasim**, K. Smith (2010) A similarity metric for quantifying system performance at pedestrian detection. *European Conference on Human Centered Design for Intelligent Transport Systems*.

Biography

Askar Obulkasim (1982, Kashgar) studied statistics, data mining, and bioinformatics. He received his M.A. in Information Management from the Xinjiang University in 2006 with highest distinction (summa cum laude). After working one year as a Data Analyst, in 2007 he began his master's studies in Statistics specializing in Machine Learning and Data Mining at the Department of Computer and Information Science (IDA), Linköping University (LiU), Sweden. In the final year of his studies, he has been given an opportunity to work in an EU project (FNIR) under the supervision of Prof. Anders Grimvall and Prof. Kip Smith for his master's thesis, and received his M.Sc. in Statistics in 2009. In the same year, he started his Ph.D. research in Bioinformatics at the Department of Epidemiology and Biostatistics, VU university medical center under the supervision of Prof. Mark van de Wiel and Prof. Gerrit Meijer, which resulted in underlaying thesis.

Currently, he is working part-time as a postdoctoral researcher in statistics for integrative bioinformatics at the Department of Paediatric Oncology, Erasmus University Medical Center, and part-time as a data scientist at The Hyve.